FEATURE EXTRACTION FROM MAMMOGRAPHIC MASS SHAPES AND DEVELOPMENT OF A MAMMOGRAM DATABASE

G. ERTAŞ[1], H. Ö. GÜLÇÜR[1], E. ARIBAL[2], A. SEMİZ[2]

[1]Institute of Biomedical Engineering, Bogazici University, Istanbul, Turkey
[2]Medical Hospital of Marmara University, Istanbul, Turkey

Abstract- **Breast cancer is one of the most common malignancies in women and a rare malignancy in men. Women who are diagnosed at an early stage can survive this often deadly disease. Mammography provides the best screening modality for detecting early breast cancer, even before a lesion is palpable. Because of the malignant mass pathology, the shape of the mammographic mass can be used to discriminate between malignant and benign masses. In this study the use of shape features to classify breast masses has been investigated and a classification scheme has been developed to classify masses as either benign or malignant. A mammogram database designed to store the images of the masses, calculated shape descriptor parameters and some additional data, such as patient history, category of the mass and biopsy report if performed which are required in BI-RADS is also introduced A touch on memory system has been used as a tool that permits access to the electronic patient record in the mammogram database. The software is written in Delphi and runs on Windows operation systems.**

Keywords **- Breast cancer, Bayesian classifier, mammography database, touch on memories.**

I. INTRODUCTION

Breast cancer is one of the most common malignancies in women and a rare malignancy in men. The earlier that a breast malignancy is detected the better the chance for a cure. It has been widely reported that breast cancer has become the second leading cause of cancer death among women. Over a lifetime, one in nine women risk contracting breast cancer. However, the good news is that women who are diagnosed at an early stage can survive this often deadly disease. Mammography provides the best screening modality for detecting early breast cancer, even before a lesion is palpable.

Approximately 80 to 85% of localized breast cancers are diagnosed by the mammographic appearances of the tumor [1]. Because of the malignant mass pathology, the shape of the mass can be used to discriminate between malignant and benign masses [2]. The application of shape analysis has been applied to computerized mammographic methods. Magnin *et al.* and Davies and Dance, and Shen *et al.* utilized shape descriptors which included various measures of area, compactness, eccentricity, and convexity [3]. However their methods were not applied to the classification of masses. The use of shape features to classify breast masses has also been investigated.

In this study geometric parameters such as area, perimeter, circularity, normalized circularity, radial distance mean and standard deviation, area ratio, orientation, eccentricity, moment invariants and Fourier descriptors up to 10, are measured. The process starts with a segmentation phase, in which an expert radiologist segments the mammographic mass shapes within the mammographic database set. These pre-segmented mammographic mass shapes are then processed by a mass boundary detection algorithm to obtain the descriptive geometric parameters. A carefully designed classification scheme is used in the final step to classify masses as benign or malign.

The results show that normalized circulatory area and the Fourier coefficient A_0 can be used successfully for feature extraction. The software developed utilizes this finding in the automatic classification of the masses.

The next part of this study includes a mammogram database design to store the images of the masses, calculated shape descriptor parameters and some additional data, such as patient morphologic data, patient medical history, category of the mass and biopsy report if performed which are required in ACR's Breast Imaging Reporting and Data System. The developed database is designed to be an Open Database Connectivity compliant relational database to support some future uses, such as screening the growth of suspicious masses, telemedical service support for sharing mass information and for facilitating statistical data analysis. The electronic record privacy of the patient is guarantied by employing a Touch Memory system. The block diagram of the developed system is shown in Fig. 1.

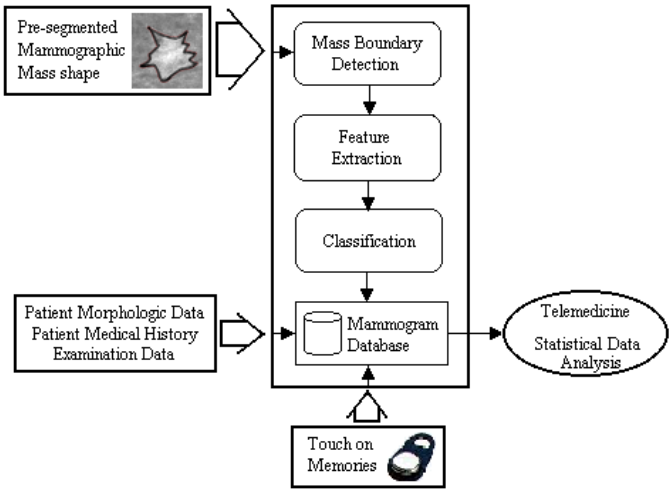

Fig. 1. Block diagram of the developed system

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Feature Extraction From Mammographic Mass Shapes and Development of A Mammogram Database	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Institute of Biomedical Engineering Bogazici Univesity Istanbul, Turkey	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom. , The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

II. Mass Shape Feature Extraction

Several qualitative and quantitative techniques have been developed for characterizing the shape of masses in an image. These techniques are useful for classifying masses in a pattern recognition system and for symbolically describing masses in an image understanding system [4].

Because of the malignant mass pathology, the shape of the mass can be used to discriminate between malignant and benign masses. In this study geometric parameters such as area, perimeter, circularity, normalized circularity, radial distance mean and standard deviation, area ratio, orientation, eccentricity, moment invariants and Fourier descriptors up to 10, are measured.

In the first step, the mammographic mass shapes, in the training data set consisting of a total of 60 cases including 30 benign and 30 malignant cases, are segmented by an expert radiologist. Geometric parameters of the pre-segmented mammographic mass shapes within this training set are then automatically computed using a software specially developed for this purpose. A careful examination of the database and experimentation with various classification techniques show that, normalized circulatory area and Fourier coefficients can be used successfully to classify masses as benign or malignant. This has been described in the second section.

A. Area

The most trivial shape parameter is the area of a mass. In a digital binary image, the area is given by the number of pixels that belong to the mass. In the matrix or pixel list representation of the mass, area computing simply means counting the number of pixels. On the other hand, the chain code of the mass can be used to calculate the area.

B. Perimeter and Normalized Circularity

The perimeter is another geometrical parameter, which can easily be obtained from the chain code of the mass boundary. It is only needed to count the length of the chain code and take into consideration that steps in diagonal directions are by a factor of $\sqrt{2}$ longer. The perimeter p is then given by an 8-neighborhood chain code, given in (1).

$$p = n_e + \sqrt{2} n_o \tag{1}$$

where n_e and n_o are the number of even and odd chain code steps, respectively [5].

C. Normalized Circularity

Area and perimeter are two parameters which describe the size of a mass in one or the other way. In order to compare masses which are observed from different distances, it is important to use shape parameters which do not depend on the size of the mass on the image plane. The normalized circularity c_N is one of the simplest parameters of this kind. It is defined in (2).

$$c_N = 1 - \frac{4\pi A}{p^2} \tag{2}$$

where A is the area and p is the perimeter of the mass. The normalized circularity equals zero for circles and tends to unity for complex shapes [6].

C. Fourier descriptors and Fourier coefficient A_0

Fourier coefficient A_0 can be computed in a way similar to radial distance. The radial distance is measured by first detecting the centroid of the mass. The Euclidean distance from the centroid to the edge is then measured for the entire boundary. An arbitrary starting point is chosen and the boundary is followed clockwise. The radial distance is computed using equation (3):

$$d(i) = \sqrt{(x(i) - x_c)^2 + (y(i) - y_c)^2} \tag{3}$$

where (x_c, y_c) are the coordinates of the centroid, $(x(i), y(i))$ are the coordinates of the boundary pixel at the i-th location. As in the radial distance computation, Euclidean distance is measured but the starting point is chosen due to the orientation of the shape and the boundary is followed clockwise by a pre selected degree increment to compute A_0 as shown in Fig. 2. In this study, 1 degree increment is used and thus N = 360.

$$A_0 = \frac{1}{N} \sum_{i=1}^{360} d(i) \tag{4}$$

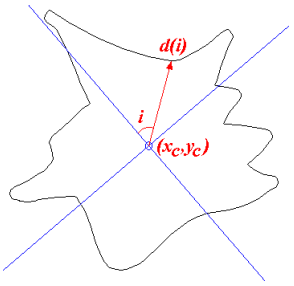

Fig. 2. Computation of A_0 for a malignant mass.

III. Classification Scheme

Pattern recognition applications come in many forms. In some instances, there is an underlying and quantifiable statistical basis for the generation of patterns. In other instances, the underlying structure of the pattern provides the information fundamental for pattern recognition. In still others, neither of the above cases hold true, but designers are

able to develop and 'train' an architecture to correctly associate input patterns with desired responses [7].

Statistical pattern recognition assumes a statistical basis for classification of algorithms. A set of characteristic measurements, denoted features, are extracted from the input data and are used to assign each feature vector to one of *c* classes. Features are assumed generated by a state of nature, and therefore the underlying model is of a state of nature or class-conditioned set of probabilities and/or probability density functions. Bayes theorem is such kind.

The Bayes decision criterion employs a systematic procedure of assigning a cost to each correct and incorrect decision and then minimizing the average cost. In this study the costs are selected specifically to hit more malignant cases. A decision tree, consisting of the Bayesian classifiers as shown in Fig. 3, is implemented to classify both of the two cases as errorless as possible. During the selection of parameters in Bayesian classifiers ROC analysis is used.

ROC analysis estimates a curve, which describes the inherent tradeoff between sensitivity and specificity of a diagnostic test [8]. Each point on the ROC curve is associated with a specific diagnostic criterion. This point will vary among observers because their diagnostic criteria will vary even when their ROC curves are the same. The area under the curve gives the accuracy of the test [9]. The parameters are selected in according to their areas under the ROC curves.

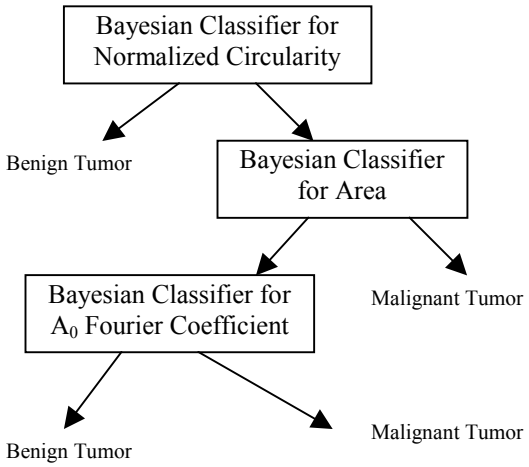

Fig. 3. The decision tree for classification

The developed computer software has been redesigned to implement the decision tree. A total number of 25 biopsy-performed mammographic mass shapes have been analyzed by the software and decisions have been made listed in Table 1.

TABLE 1
DECISIONS OF THE SOFTWARE

		Biopsy Result	
		Malignant	Benign
Decision	Malignant	9	1
	Benign	1	14

IV. MAMMOGRAM DATABASE

Database systems store information in every conceivable health care environment. From large tracking databases such as hospital information systems to a patient's examination report, database systems store and distribute the data that are depended on. Today's generation of powerful, inexpensive workstation computers enables programmers to design software that maintains and distributes data quickly and inexpensively.

A. Relational Database Management System

Codd's idea for a Relational Database Management system, shortened as RDBMS, uses the mathematical concepts of relational algebra to break down data into sets and related common subsets. Because information can naturally be grouped into distinct sets, Dr. Codd organized the database system around this concept. Under the relational model, data is separated into sets that resemble a table structure. This table structure consists of individual data elements called columns or fields. A single set of a group of fields is known as a record or row [10]. Microsoft Access is a PC-only database product that contains many of the features of a relational database management system and this is the one reason why it has been chosen for the mammogram database.

B. Open Database Connectivity

The unique feature of Open Database Connectivity, shortened as ODBC, is that none of its functions are database-vendor specific. ODBC allows to access databases from remote clients over Intranet and the most popular protocol used is TCP/IP [10]. The second reason for choosing Microsoft Access is the ODBC compatibility which enables internet access for telemedical use.

C. Structured Query Language

Structured query language, named as SQL, evolved to service the concepts of the relational database model. SQL is a nonprocedural language, in contrast to the procedural or third-generation languages such as COBOL and C. Nonprocedural means what rather than how. SQL describes what data to retrieve, delete, or insert, rather than how to perform the operation and this property allows SQL to find an application area in the developed database.

D. Tables in The Mammogram Database

The most important decision for a database designer, after the hardware platform and the RDBMS have been chosen, is the structure of the tables. Decisions made at this stage of the design can affect performance and programming later during the development process.

The developed mammogram database consists of mainly 5, total 9 numbers of tables. The tables named as *patientfile*, *reports*, *findings, biopsies* and *doctors* are mainly used to store electronic patient record. In addition to this, the tables labeled as *assessments*, *recommendations*, *reasons* and *biopsytech* are designed for the coding system of the software. The table's relationships are as shown in Fig. 4.

1) *Patient File Table:* This table stores the morphologic data of the patient including patient's ID, photo and patient medical history related to breast cancer such as menopause and menarche ages, births before and after age 30, the prior breast cancers, ovarian and endothelial cancers are seen or not, patient' s mother and sister have the symptoms of breast cancer or not.

2) *Reports Table:* The examination date, the name of the doctor who examined, breast composition and the assessment, recommendation, examination reason codes are stored in this table. While the relationship with the patient file table is obtained via patient ID, another relationship is enabled via mass findings ID.

3) *Findings Table:* Mammographic mass finding data, including the image and the location of the mass; normalized circularity, area and the Fourier coefficient A_0 parameters of the mass, software decision and the ID of the biopsy, if performed, are the fields of the findings table.

4) *Biopsies Table:* The biopsy data, consisting of the used technique, classification and the pathological size of the mass, are stored in this table.

E. Coding System

The coding system is defined with the assessments, recommendations, reasons and biopsytech tables. During the design of these tables, much attention is focused on the American College of Radiology's Breast Imaging Reporting and Data System shortened as BI-RADS since it contains a guide to standard coding.

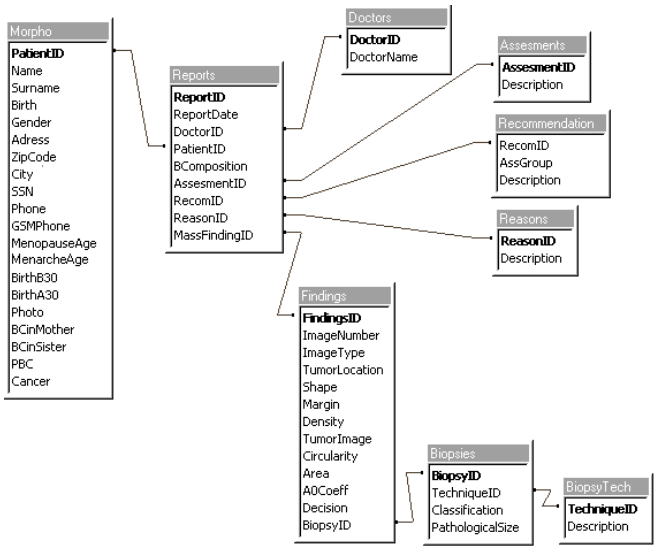

Fig. 4. Relationships of the tables in mammogram database

F. Patient Privacy

Touch Memory (TOM) buttons are electronic memory chips contained in small, water-resistant, stainless steel canisters. All Touch Memory buttons contain a unique ID number that is unalterable and identifies each button [11]. The unique ID number in TOM is used as a tool that permits access to an electronic patient record in the mammogram database via the developed computer software. DS9097 COM port adapter is employed as a TOM reader and connected to the computer via an RS232 serial port.

V. DISCUSSION

Mammography provides the best screening modality for detecting early breast cancer, even before a lesion is palpable. Because of the malignant mass pathology, the shape of the mass can be used to discriminate between malignant and benign masses.

This study defines a classification scheme based on the geometrical parameters of mammographic mass shapes and a BI-RADS compatible mammogram database structure. With respect to these definitions, a 32-bit computer software is developed for clinical use. The software makes decisions in a few seconds. In addition to this, it has a user friendly graphical interface and runs on windows operation systems.

REFERENCES

[1] R. L. Egan, *Breast Imaging: Diagnosis and Morphology of Breast Diseases*, Philadelphia, PA: Saunders,1988.

[2] S. A. Feig and R. McLelland, *Breast Carcinoma: Current Diagnosis and Treatment*, New York: Masson, 1983.

[3] L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," IEEE Trans. Med. Imag., vol. 12, pp. 263-274, June 1994.

[4] W. Pratt, *Digital Image Processing*, pp:629, 1991.

[5] B. Jahne, *Digital image processing : concepts, algorithms, and scientific applications,* New York : Springer-Verlag, 1993.

[6] I. Pitas, *Digital image processing algorithms*, New York : Prentice Hall, 1993.

[7] R. Schalkoff, *Pattern Recognition: Statistical, structural and neural approaches*, John Wiley & Sons Inc., 1992.

[8] PF. Griner, RJ. Mayewski, AI. Mushlin, P. Greenland, "Selection and Interpretation of Diagnostic Tests and Procedures", *Annals of Internal Medicine,* 94, 555-600, 1981.

[9] MH. Zweig, G. Campbell, "Receiver-operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine", *Clinical Chemistry,* 39, 561-577, 1993.

[10] M. Spenik, *Web Database Developer's Guide with Visual Basic 5*, Sams.net Publishing, 1997.

[11] Automatic Identification Data Book, Dallas Semiconductor Co., 1996.